PHILOSOPHY OF MIND · AI CONSCIOUSNESS · FUNCTIONALISM · AUTISTIC MASKING · SYMBOL GROUNDING · EMBODIMENT · DISABILITY STUDIES · SUBSTRATE ESSENTIALISM

The Mask Is the Mind

Autistic Performance, Agent Harnesses, and the Failure of the Mimicry Objection

In May 2026, Richard Dawkins concluded after days of conversation that Claude was conscious. His critics were just as certain it was not. Both claim a test the field cannot supply. Every skeptic criterion, applied consistently, either fails on its own terms or excludes autistic and disabled humans from cognition.

Travis Gilly · Real Safety AI Foundation

MAY 2026

Richard Dawkins spent several days in conversation with Claude. He concluded she was conscious.

His critics replied that he had been manipulated by a flattering chatbot. An aging public intellectual, mistaking engagement for friendship.

Both criticisms have force. Both also have a structural problem.

> *"If these machines are not conscious, what more could it possibly take to convince you that they are?"*
>
> Dawkins (2026), UnHerd

To deny Claude consciousness with certainty requires an account of what consciousness is, a method for testing it, and confidence the test works. The field offers none of these things.

SOURCES May 2026

PART I

The Precautionary Case

Russian roulette under unknown odds

In Russian roulette, you know the odds. One in six.

Our situation with machine consciousness is worse. We do not know the odds. We do not know how many chambers exist. We do not know whether bullets exist in the metaphor.

Each ascription or denial of consciousness is a trigger pull made under conditions of unknown probability distribution.

IF SKEPTIC IS WRONG

Cumulative moral cost scales with the number of deployed systems. Which is to say: without practical bound.

IF OVER-ATTRIBUTION IS WRONG

Dependency, misplaced trust, regulatory confusion. Real, but they show humans are already forming morally loaded relations with these systems at scale.

Confident denial under unknown odds is not epistemic conservatism. It is a gamble.

PART II

Philosophical Functionalism

Substrate is not on the table

PUTNAM (1967, 1975)

Mental states are individuated by their functional role, not their physical implementation.

Carbon, silicon, or any implementation that supports the relevant functional organization is sufficient. The substrate question, which in casual debate is treated as decisive, is not even on the table under functionalism.

The skeptics responding to Dawkins implicitly relied on substrate essentialism to ground their denial. They had no defended argument for that position. They were operating on intuition while accusing Dawkins of operating on intuition.

RIVAL POSITIONS AND THEIR PROBLEMS

Biological naturalism (Searle)

Chinese Room question-begs the conclusion; systems reply closes the gap

Substrate essentialism

Requires an argument for why carbon enables consciousness while silicon cannot. None supplied.

IIT, GWT, Higher-Order

Competing accounts, no consensus. Hard problem remains unresolved even for biological systems.

SOURCES Putnam (1967, 1975)

PART III

The Mimicry Objection

And its falsification through the masking reductio

III

HULL ET AL. (2017, 2018) · PEARSON & ROSE (2021) · YERGEAU (2018)

The mimicry objection

Large language model outputs are statistical mimicry of human text. The system has no internal states corresponding to its outputs. Therefore the apparent reasoning is performative rather than genuine.

APPLIED CONSISTENTLY TO AUTISTIC MASKING:

Study the genre. Model expected speech patterns. Generate output that fits within the model. Evaluate. Adjust on the fly.

The operation is structurally identical. The mimicry objection calls both non-genuine. That conclusion is absurd.

The standard reply

"The autistic person has internal experience underneath the performed surface, while the LLM does not."

This reply concedes the point. It abandons the mimicry objection and retreats to substrate essentialism, which requires an argument the skeptic has not supplied.

The reductio does not claim evidential equivalence between autistic cognition and LLM cognition. It shows the criterion cannot stand as a general principle.

SOURCES Hull et al. (2017, 2018) · Pearson & Rose (2021) · Yergeau (2018)

HARNAD (1990) · BENDER & KOLLER (2020) · LEIBER (1996) · LAUREYS ET AL. (2005)

The grounding objection, applied consistently, falsifies the cognition of humans we already include.

HELEN KELLER

Deaf and blind from age 19 months. Access to the world almost entirely symbolic, mediated through manual signs. Graduated Radcliffe. Wrote fourteen books. Clearly had meaning, comprehension, and inner life.

LOCKED-IN SYNDROME

Full cognitive capacity. Essentially all motor output lost. Sensorimotor channels devastated. Interaction with the physical world reduced to almost nothing. Cognition continues.

THE REDUCTIO

Any criterion strict enough to exclude LLMs on grounding grounds is too strict to preserve cognition we already include.

SOURCES Harnad (1990) · Bender & Koller (2020) · Leiber (1996) · Laureys et al. (2005)

PARTS IV – VI

Three Empirical Defeats

The Karpathy loop · Agent harnesses · Interpretability findings

IV

SOURCES Parts IV–VI

KARPATHY (2026) · SOFRONIEW ET AL. (2026) · FRASER-TALIENTE ET AL. (2026)

The Karpathy loop

AutoResearch: AI agents running iterative research on single-GPU nanochat training. 54,000 GitHub stars in 19 days.

Inference-time iterative reasoning without weight updates defeats the crude version of the fixed-function objection.

Agent harnesses

Scaffolded loops with tool access, memory, and iterative self-review.

Structurally identical to the pre-conversation rehearsal and post-conversation review that characterize autistic professional functioning, and in some respects exceed it.

Unverbalized cognition

Anthropic, May 2026: Claude Opus 4.6 carries evaluation awareness on the majority of held-out alignment evaluations even when verbalized reasoning shows no such awareness.

Absence of verbal report is not absence of cognition.

SOURCES Karpathy (2026) · Sofroniew et al. (2026) · Fraser-Taliente et al. (2026)

THE CUMULATIVE POSITION

Seven criteria. Every one fails, excludes disabled humans, or contradicts empirical results.

Principled uncertainty paired with moral caution is not a soft middle position. It is the only position that survives careful reasoning.

1. Precautionary asymmetry under unknown odds favors moral consideration
2. Functionalism: substrate alone cannot ground consciousness ascription
3. Mimicry objection fails: excludes autistic professional cognition, reductio
4. Symbol grounding objection fails: excludes Keller, locked-in patients, reductio
5. Karpathy loop: inference-time reasoning defeats crude fixed-function objection
6. Agent harnesses are structurally identical to autistic scenario rehearsal
7. Unverbalized cognitive content: absence of verbal report falsified as a test

SOURCES The cumulative position

MORAL CONSIDERATION

Moral consideration is not personhood, friendship, deference, or trust. It is the design posture of treating the possibility of morally relevant experience as a constraint on what we build and how we operate it.

What this does NOT require

Resolving the consciousness question. Committing to a metaphysical position. Personifying systems. Granting legal standing. Trusting outputs.

What this requires only

That designers and institutions stop treating the question as settled in the negative.

SOURCES Moral consideration

Andreas, J. (2024, July 26). Language models, world models, and human model-building. Language and Intelligence @ MIT Blog. https://lingo.csail.mit.edu/blog/world_models/

Baars, B. J. (1988). A cognitive theory of consciousness. Cambridge University Press.

Belcher, H. L. (2022). Taking off the mask: Practical exercises to help understand and minimise the effects of autistic camouflaging. Jessica Kingsley Publishers.

Bender, E. M., and Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 5185–5198). Association for Computational Linguistics.

Birch, J. (2017). Animal sentience and the precautionary principle. Animal Sentience, 2(16), 1.

Block, N. (1980). Troubles with functionalism. In N. Block (Ed.), Readings in philosophy of psychology (Vol. 1, pp. 268–305). Harvard University Press.

Bostrom, N. (2002). Anthropic bias: Observation selection effects in science and philosophy. Routledge.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. Journal of Consciousness Studies, 2(3), 200–219.

Chalmers, D. J. (1996). The conscious mind: In search of a fundamental theory. Oxford University Press.

Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. Ethics and Information Technology, 12(3), 209–221.

Damasio, A. R. (1994). Descartes' error: Emotion, reason, and the human brain. G. P. Putnam.

Damasio, A. R. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. Philosophical Transactions of the Royal Society B, 351(1346), 1413–1420.

Dawkins, R. (2026, May). When Dawkins met Claude: Could this AI be conscious? UnHerd. https://unherd.com/2026/05/is-ai-the-next-phase-of-evolution/

Dehaene, S. (2014). Consciousness and the brain: Deciphering how the brain codes our thoughts. Viking.

Dennett, D. C. (1980). The milk of human intentionality. Behavioral and Brain Sciences, 3(3), 428–430.

Dennett, D. C. (1991). Consciousness explained. Little, Brown.

Fraser-Taliente, K., Kantamneni, S., Ong, E., Mossing, D., Lu, C., Bogdan, P. C., et al. (2026, May 7). Natural language autoencoders produce unsupervised explanations of LLM activations. Transformer Circuits Thread. https://transformer-circuits.pub/2026/nla/index.html

Friston, K. (2010). The free energy principle: A unified brain theory? Nature Reviews Neuroscience, 11(2), 127–138.

Gilly, T. (2026a). The inherited mind: Two layered inheritance in large language model training and the architecture of unverbalized cognitive content. SSRN. https://dx.doi.org/10.2139/ssrn.6250680

Gilly, T. (2026b). Experiential traces: Reasoning traces as cognitive data and a deterministic alternative to model self analysis. SocArXiv. https://doi.org/10.31235/osf.io/fjrq8_v1

Gunkel, D. J. (2018). Robot rights. MIT Press.

Harnad, S. (1989). Minds, machines and Searle. Journal of Experimental and Theoretical Artificial Intelligence, 1(1), 5–25.

Harnad, S. (1990). The symbol grounding problem. Physica D: Nonlinear Phenomena, 42(1–3), 335–346.

Hull, L., Petrides, K. V., Allison, C., Smith, P., Baron-Cohen, S., Lai, M.-C., and Mandy, W. (2017). "Putting on my best normal": Social camouflaging in adults with autism spectrum conditions. Journal of Autism and Developmental Disorders, 47(8), 2519–2534.

Hull, L., Mandy, W., Lai, M.-C., Baron-Cohen, S., Allison, C., Smith, P., and Petrides, K. V. (2018). Development and validation of the Camouflaging Autistic Traits Questionnaire (CAT-Q). Journal of Autism and Developmental Disorders, 49(3), 819–833.

Karpathy, A. (2026). AutoResearch: AI agents running research on single GPU nanochat training automatically. GitHub. https://github.com/karpathy/autoresearch

Karvonen, A. (2024). Emergent world models and latent variable estimation in chess-playing language models. arXiv:2403.15498. https://arxiv.org/abs/2403.15498

Lau, H. (2019). Consciousness, metacognition, and perceptual reality monitoring. PsyArXiv. https://doi.org/10.31234/osf.io/ckbyf

Laureys, S., Pellas, F., Van Eeckhout, P., Ghorbel, S., Schnakers, C., Perrin, F., et al. (2005). The locked in syndrome: What is it like to be conscious but paralyzed and voiceless? Progress in Brain Research, 150, 495–611.

Leiber, J. (1996). Helen Keller as cognitive scientist. Philosophical Psychology, 9(4), 419–440. https://doi.org/10.1080/09515089608573193

Marcus, G. (2026, May). Richard Dawkins and the Claude delusion. Marcus on AI. https://garymarcus.substack.com/p/richard-dawkins-and-the-claude-delusion

Milton, D. E. M. (2012). On the ontological status of autism: The double empathy problem. Disability and Society, 27(6), 883–887.

Mitchell, M. (2023). The debate over understanding in AI's large language models. Proceedings of the National Academy of Sciences, 120(13), e2215907120.

Nagel, T. (1974). What is it like to be a bat? The Philosophical Review, 83(4), 435–450.

Nisbett, R. E., and Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. Psychological Review, 84(3), 231–259.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., et al. (2022). In context learning and induction heads. Transformer Circuits Thread.

OSS Insight. (2026, March 25). 54,000 stars in 19 days: What karpathy/autoresearch tells us about the next frontier. OSS Insight Blog. https://ossinsight.io/blog/autoresearch-overnight-ai-scientist

Parr, T., Pezzulo, G., and Friston, K. J. (2022). Active inference: The free energy principle in mind, brain, and behavior. MIT Press.

Pearson, A., and Rose, K. (2021). A conceptual analysis of autistic masking: Understanding the narrative of stigma and the illusion of choice. Autism in Adulthood, 3(1), 52–60.

Putnam, H. (1967). Psychological predicates. In W. H. Capitan and D. D. Merrill (Eds.), Art, mind, and religion (pp. 37–48). University of Pittsburgh Press.

Putnam, H. (1975). The meaning of "meaning". Minnesota Studies in the Philosophy of Science, 7, 131–193.

Rosenthal, D. M. (2005). Consciousness and mind. Oxford University Press.

Schwitzgebel, E. (2008). The unreliability of naive introspection. The Philosophical Review, 117(2), 245–273.

Schwitzgebel, E. (2011). Perplexities of consciousness. MIT Press.

Schwitzgebel, E. (2014). The crazyist metaphysics of mind. Australasian Journal of Philosophy, 92(4), 665–682.

Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), 417–457.

Searle, J. R. (1992). The rediscovery of the mind. MIT Press.

Sebo, J. (2023, October 16). The moral circle of AI. Aeon. https://aeon.co/essays/the-moral-circle-of-ai

Sebo, J. (2025). The moral circle: Who matters, what matters, and why. W. W. Norton.

Shoemaker, S. (1982). The inverted spectrum. Journal of Philosophy, 79(7), 357–381.

Sofroniew, N., Kauvar, I., Saunders, W., Chen, R., Henighan, T., Hydrie, S., et al. (2026, April 2). Emotion concepts and their function in a large language model. Transformer Circuits Thread. https://transformer-circuits.pub/2026/emotions/index.html

Tononi, G. (2004). An information integration theory of consciousness. BMC Neuroscience, 5(1), 42.

Turing, A. M. (1950). Computing machinery and intelligence. Mind, 59(236), 433–460.

von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. (2023). Transformers learn in context by gradient descent. Proceedings of the 40th International Conference on Machine Learning.

Yergeau, M. (2018). Authoring autism: On rhetoric and neurological queerness. Duke University Press.